JLIS.it

Transforming and enhancing metadata for enduser discovery: a case study

Edward M. Corrado, Rachel Jaffe

1 Introduction

This paper describes the process developed by Binghamton University Libraries to extract embedded metadata from digital photographs and transform this metadata into descriptive metadata for use in the Libraries' digital preservation system.

In 2011, the Libraries implemented the Rosetta digital preservation system to preserve digitized and born-digital materials. The Libraries' are working to preserve a variety of digital objects including items located in the Special Collections and Archives, as well as faculty and student research, and materials produced by the University's Office of Communications and Marketing. At the same time, the Libraries' implemented the Primo discovery tool to bring together the digital collections in Rosetta, bibliographic holdings from our integrated library system, and data from other sources. The Libraries' have been working to preserve a variety of digital collections including items located in the Special Collections and Archives, as well as faculty and student research, and materials produced by

JLIS.it. Vol. 5, n. 2 (Luglio/July 2014).
DOI: 10.4403/jlis.it-10069

the University's Office of Communications and Marketing.

Each object that is deposited in the digital preservation system is accompanied by descriptive and technical metadata. Binghamton University Libraries uses the Dublin Core Metadata Element Set (DCMES) to record this metadata. We have found that DCMES is flexible and accommodates most needs and use over fifty of the DCMES elements and qualifiers. For large photographic collections, the Libraries creates most of the DCMES metadata by transforming and enhancing metadata that has been embedded into the photographs.

2 The Project

2.1 Scope

The Libraries are currently working with the University's photographer to preserve and provide access to over 350,000 digital images. Most of these images depict events (including athletic events, Homecoming, and Commencement) that are of historical and immediate social value to the University community. The photographer's images are used widely in marketing and outreach materials, and on the University's website. Owing to volume of photographs, as well as to budgetary and other constraints, it is not possible to have library staff inspect and create a complete descriptive metadata record for each photograph, so the Libraries needed to explore different options.

Each of photographer's images contains embedded technical metadata (file format, date and time, etc.) and additionally, many of the files also include basic descriptive information supplied by the photographer, such as name, keywords and description. Using this basic metadata as a starting point, librarians were able to

create an automated process to reformat and enhance the available descriptive information, crosswalk it to the DCMES, and map the photographer's keywords to controlled subject terms.

2.2 Literature Review

2.2.1 Digital preservation

The Digital Preservation Coalition describes digital preservation as a "series of managed activities necessary to ensure continued access to digital materials for as long as necessary [...] beyond the limits of media failure or technological change»(Digital Preservation Coalition). While digital preservation involves a great deal of technology it is not merely a technical problem. As the National Library of New Zealand's Steve Knight has commented, «Digital preservation requires interaction with all the organisation's processes and procedures». (Knight) Digital Preservation cannot be «accomplished in isolation by staff in remote parts of the institution unfamiliar with the mission, goals, users, content, and culture of the organization»(Corrado and Moulaison). Because digital objects can only be read with software, having information, or metadata, about digital objects «is a key factor for ensuring the long-term access of digital resources»(Gelaw, Hastings, and Hartman).

2.2.2 Metadata

Metadata are the elements used to describe resources for the purposes of discovery, rights management and preservation. According to the National Information Standards Organization (NISO), «Metadata is often called data about data or information about information»(National Information Standards Organization (NISO)). This is a common but not the most useful definition. Metadata assist

users in identifying, authenticating and contextualizing data, data sets and other digital resources as well as to describe the structural relationships within and between these materials. Metadata are used to define permissions — access rights, sharing, re-use and re-distribution policies — as well as the technical requirements for viewing, accessing or preserving born digital and/or digitized objects.

Within library and information science literature, metadata for digital collections is often divided into categories based on function. Some authors present three main types of metadata, NISO and Miller (*Metadata for digital collections*) list administrative (inclusive of rights, technical and preservation metadata), descriptive and structural metadata; Gilliland-Swetland (*Introduction to metadata: Setting the stage*) and the Getty's Introduction to Metadata (Getty Research Institute) define five categories and others give four (Corrado and Moulaison, p. 113). Despite this variation, overlap and indecision about which kind of metadata belongs in which category, as Corrado and Moulaison observe, «In truth, it really does not matter which category metadata is assigned as long as necessary metadata is provided, consistently created and input, and accessible through the system»(, p. 114).

In this project, our primary concern was capturing and creating descriptive metadata. Descriptive metadata records the attributes of a resource; it provides both intellectual access to content and access points by which users can discover digital materials. Of the different schemas or sets of fields that can be used to make up a metadata record, the DCMES is perhaps the most widely used. DCMES includes 15 core elements, plus a number of refinements (Dublin Core Metadata Initiative). Developed to describe born digital documents, DCMES was designed to sufficiently record the most basic units of information needed to facilitate basic enduser tasks (Coyle).

Despite its simplicity and extreme flexibility, use of standards like DCMES and controlled vocabularies like Library of Congress Subject Headings (LCSH) or Dublin Core Metadata Initiative (DCMI) Type Vocabulary ensures the quality of metadata across collections and institutions.

2.2.3 Strategies for dealing with big digital photo collections

While big data is widely discussed, managing and describing big photos is not. Large photographic collections, whether digitized or born digital, present a unique challenge to librarians. Often undescribed entirely or undescribed at the item level, librarians must grapple first with the volume of photographs and secondly with the question of how to re-contextualize, generate access points, and determine the origins and technical specifications of these images. Or more broadly, how does one best create full metadata records for these objects?

Much of the literature focuses on the issue of metadata creation — metadata workflows and their efficiency, evaluating the quality of the metadata produced and finding a balance between the two. The need to streamline metadata workflows and shift the burden of work of specialists is an interest or concern noted by many. Of late, consensus has risen around the principle of «more product, less process»developed and forwarded by Mark A. Greene and Dennis Meissner with respect to physical collections (Greene and Meissner). Greene and Meissner encourage archivists to do «the least we can do to get the job done in a way that is adequate to user needs, now and in the future»(, p. 240). In response many institutions are exploring and employing strategies for automating at least some steps within metadata workflows. Among the solutions and tools being used are: crowdsourcing (Raymond); education, i.e. encouraging image creators to embed descriptive and rights metadata (Keough and

Wolfe); facial recognition software (Banerjee and Anderson); and the extraction and reuse of embedded metadata (Walsh, "Repurposing embedded image metadata for DSpace batch loading (XMP to CSV to DSpace Dublin Core)"; "Automated reuse of embedded image metadata for the Knowledge Bank").

2.2.4 Workflow

While our focus is on how large sets of images are processed within the Libraries, it is important to note that the workflow actually begins before the images are received. After shooting an event, the photographer's first task is to review and select his best images. Of the photographs he selects, the photographer uses photo editing software, in this case Photoshop, to update the file names and adds some basic descriptive and rights information to the photographs. Most of this information is assigned to all or many of the photographs in a given set (i.e. collection or sub-collection); however, when there is an especially fine photograph, he will supply additional keywords and/or perhaps a unique description. Although some of descriptive metadata that we wanted to use was already in DCMES, the data that was entered into each field did not conform the Libraries' metadata best practices. Consequently, beyond the extraction of the embedded metadata, further processing was still required.

After receiving the image files from the photographer in TIFF format, the first step is to review the embedded metadata and determine which of the fields should be retained and how they will map to the Libraries' local DCMES set. Specific information or fields not included in the photographer's embedded metadata must also be identified in order to add them later.

After making these determinations, librarians need to consider how they want the data in each of the metadata fields to appear and how to reformat the data if necessary. From the embedded metadata,

librarians are able to populate the DCMES Format Medium, Creator, Description, Date Created, and various subject fields. The librarians also want to add collection-specific metadata, including the collection and sub-collection names, rights, and license information. Some of this information did appear in the photographer's embedded metadata; however, owing to inconsistent formatting and other issues, it was easier to discard and recreate this information than it was to reformat it. Additionally the file names associated with each photograph are read in from the file system and used to populate the DCMES Identifier field. Lastly, the embedded metadata does not include titles for the images, so the sub-collection names are used as the title for each of the images within a given set.

After determining the metadata mappings, reformatting the data and adding new fields, the next task is to map the embedded keywords to controlled subject headings. The photographer aided this process greatly having had the foresight to apply consistent keywords. The Libraries have formalized these keywords to create a keyword-mapping table for use in describing University materials in the digital preservation system.

In order to create the keyword mapping table, librarians extracted the embedded keywords from the first set of photographs, listed them in the order of the frequency, and isolated the terms determined to be of most value to endusers in search and retrieval. These keywords are then mapped to the appropriate Library of Congress Subject Headings (LCSH), Thesaurus for Graphic Materials (TGM), and Getty Thesaurus of Geographic Names (TGN) headings. If a photograph has not been assigned keywords, generic LCSH (e.g., State University of New York at Binghamton–Pictorial works) are added to the metadata. With each new load of photographs this process is repeated and new terms are added to the keyword mapping table.

While this process initially seems like a lot of work and some may question the value of it, given the volume of photographs the Libraries are processing, it can be viewed as an investment. The librarians are able to produce accurate, consistent and complete descriptive metadata records, which can then be integrated and indexed with metadata from other sources, and ultimately made discoverable by endusers. After all, while «preservation is critical for us, so is retrieval»(Corrado and Card). No matter how well a digital preservation system preserves a file it makes no sense to put digital objects into a system if one cannot find them later.

2.2.5 Technical Details

As part of a project's initial setup, librarians created a project specific metadata form, inclusive of all the DCMES fields being used, and a DCMES mapping table within the digital preservation system. This mapping table is used to map external metadata to the appropriate fields within the system. For the University's photograph collection, the systems librarian created shell scripts to extract the basic embedded metadata from the individual photographs; a second script compares this metadata to the keyword mapping table and maps or assigns controlled subject terms. The complete DCMES metadata is then written to a Comma-Separated Value (CSV) formatted text file that can be deposited into the digital preservation system. This CSV file is then uploaded, along with the original photographs, into the digital preservation system.

When we first looked at the embedded metadata using various tools, we found that there were multiple metadata schemas in use but they did not all provide the same information. It was determined that the most useful descriptive metadata was stored either in Exif or as XMP using DCMES. In order to extract the metadata we used ExifTool an application designed «for reading, writing and editing

meta information in a wide variety of files».[1]

Although there was some DCMES metadata embedded into the photographs, this metadata still needed to be processed for various reasons so that it would conform to the Libraries' best practices. For example, the photographer name was stored in all capital letters with first name first but the Libraries best practices were to store the names with standard name capitalization and in last name, first name order. Ultimately the fields we extracted from the embedded metadata or gathered otherwise from the images included filename, file type, file mime type, date created, image size, creator, description, and keywords. The most challenging task was in figuring out how to make the automatic mapping of keywords happen. After the creator keywords were mapped to controlled vocabulary terms and the keyword mapping table was created, the systems librarian created a shell script that reads in the keywords that were extracted from the photographs and compares them to the mapping table. The script then outputs the appropriate controlled vocabulary terms. The script also deduplicates the controlled subject terms assigned to each metadata record as there are cases where multiple keywords or variations of a single keyword map to the same controlled term. Besides mapping these keywords to controlled subject fields, the original keywords in the embedded metadata are also ingested into the digital preservation system in order to provide «a fuller representation of the intellectual content of information objects and ultimately improve subject access for the users»(Zavalina). Additionally, the script also creates or transforms other metadata that is to be ingested into the digital preservation system. This includes replacing certain types of information such rights statements, reformatting dates, and moving information from one field to another as defined by the Libraries' best practices. Figure **??** on page ?? shows

[1] «ExifTool by Phil Harvey», http://www.sno.phy.queensu.ca/~phil/exiftool.

the workflow for the initial set-up of this project.

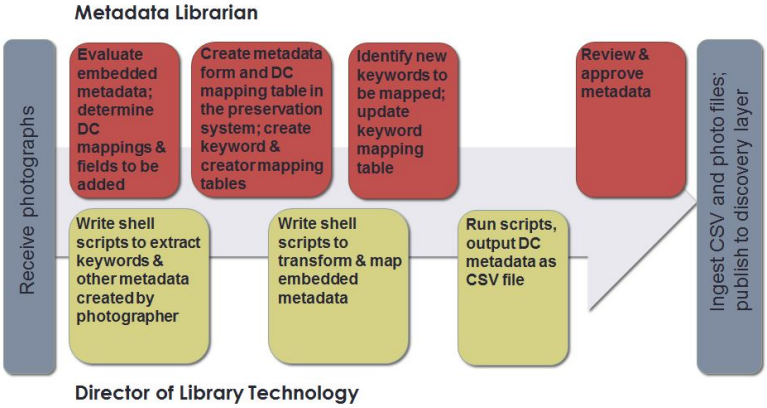

Figure 1: Workflow for Initial Set-Up

.

For subsequent sets of photos, the only steps requiring manual intervention are extracting and identifying any new keywords that might not have existed in earlier sets, updating the keyword mapping table, and finally loading the images and metadata into the digital preservation system. As shown in Figure 2 on the next page, the workflow for subsequent loads is less involved because much of the initial set-up does not need to be repeated. Once the metadata and the photographs are deposited into the digital preservation system, the metadata needs to be harvested by the Primo discovery layer software. This is done using Primo's Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) harvester that has been customized to harvest metadata from the Libraries' digital preservation system. OAI-PMH is designed to be «a low-barrier mechanism for repository interoperability»(*Open Archives Initiative Protocol for Metadata Harvesting*). In order to satisfy the need for low-barrier

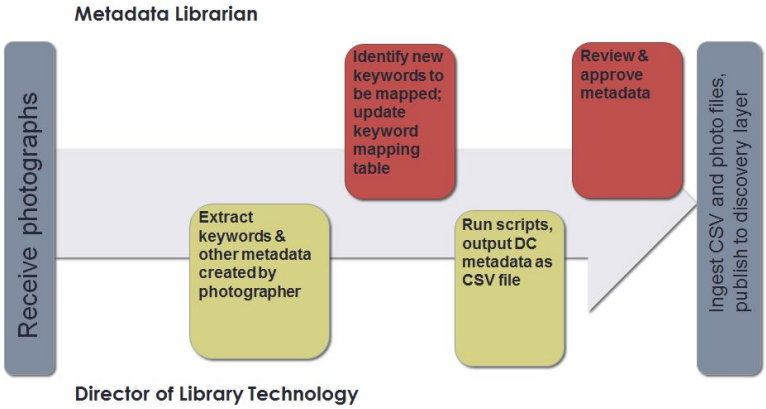

Figure 2: Workflow for Subsequent Loads

.

interoperability, OAI-PMH mandates that «repositories must be able to return records with metadata expressed in the Dublin Core format, without any qualification»(*Open Archives Initiative Protocol for Metadata Harvesting*). This creates a least common denominator problem where a great deal of specificity may be lost. Although «optionally, a repository may also disseminate other formats of metadata»(*Open Archives Initiative Protocol for Metadata Harvesting*) this does not always happen and even when it does the additional formats may not be useful to the system harvesting the metadata. By default, the Libraries' digital preservation system did not publish the metadata in qualified DCMES; therefore the OAI-PMH provider portion of the digital preservation system and Primo's OAI-PMH harvester had to be modified to include the qualified fields that librarians wanted to expose to endusers. Likewise, Primo's OAI-PMH harvester, which by default did not understand qualified DCMES had to be modified so it could properly harvest the metadata records from the digital

preservation system. Since we had control of both systems this task was not difficult but this could be a challenge in other environments where librarians don't control both ends of the system and/or for those who are attempting to federate metadata from multiple systems where «metadata quality in relation to interoperability are especially pronounced»(Park).

3 Conclusion

As this process is refined and is considered for use with other digital collections, it has become apparent that both technical and descriptive metadata can be found anywhere — it is just a matter of locating the metadata and transforming it into a format meaningful to endusers. Sometimes basic descriptive metadata is embedded within digital objects and other times it might be used as part of the file naming or directory structure. While the quality and richness of the metadata supplied may vary, as long as it has been applied in a somewhat consistent manner, it can be repurposed and enhanced in order to create more complete metadata record.

The approach described in this paper and related approaches may be useful in many settings. The key is to look holistically at a collection of digital objects to see what kind of metadata is available that can be transformed into something meaningful. While it would be ideal to catalog each item individually, it is often neither practical nor even possible due to time and funding constraints.

«The greatest danger to digital materials is that we forget the meaning of them»(Lesk). Metadata is an important aspect of digital preservation that helps ensure that we do not forget the meaning of our digital objects. This goes beyond the technical and administrative metadata that is used by digital preservation software to determine file formats and to ensure the integrity of digital objects

and includes the need for descriptive metadata. No matter how well the bits and bytes of a digital object are preserved, it is meaningless if the object cannot be discovered and retrieved. Ultimately digital preservation is for use and something that cannot be found cannot be used. As Gilliland-Swetland so eloquently put it, «Metadata is like interest — it accrues over time. To stretch the metaphor further, wise investments generate the best return on intellectual capital. Carefully designed metadata results in the best information management in the short and long-term»(Gilliland-Swetland, p. 11). While the size and scope of today's digital collections make it nearly impossible to have a cataloger or metadata librarian individual describe each individual item, «[t]he cataloging community has traditionally been a community of innovators»(Moulaison, "A New Cataloging Curriculum in a Time of Innovation: Exploring a Modular Approach to Online Delivery"). One way in which cataloging, metadata and systems librarians can innovate is by adjusting their roles. In this paper we describe one such way that this can be accomplished. By bringing traditional cataloging skills together with technology, librarians can provide enhanced access to digital objects that might not otherwise be discoverable.

References

Banerjee, Kyle and Maija Anderson. "Batch metadata assignment to archival photograph collections using facial recognition software". *The Code4Lib Journal* 21. (July 2013). http://journal.code4lib.org/articles/8486. (Cit. on p. 38).

Corrado, Edward M. and Sandy Card. *Digital Preservation Policy and Procedures*. Athens, Georgia, Paper presented at the annual meeting of the ex libris users of north america, athens, georgia, april 30 - may 3, 2013. (Cit. on p. 40).

Corrado, Edward M. and Heather Lea Moulaison. *Digital preservation for libraries, archives, and museums*. Lanham: Rowman & Littlefield, 2014. (Cit. on pp. 35, 36).

Coyle, Karen. "Understanding Metadata and Its Purpose". *Journal of academic librarianship* 2. (2005): 160–163. (Cit. on p. 36).

Digital Preservation Coalition. *Introduction: Definitions and Concepts*. 2012. http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts. (Cit. on p. 35).

Dublin Core Metadata Initiative. *Dublin Core Metadata Element Set, Version 1.1*. 2012. http://dublincore.org/documents/dces. (Cit. on p. 36).

Gelaw, Daniel, S.K. Hastings, and Cathy Hartman. "A Metadata Approach to Preservation of Digital Resources: The University of North Texas Libraries' Experience". *First Monday* 7.8. DOI: 10.5210/fm.v7i8.981. (Aug. 2002). (Cit. on p. 35).

Getty Research Institute. *Introduction to metadata*. Ed. Murtha Baca. 2nd ed. Los Angeles, CA: Getty Research Institute, 2008. (Cit. on p. 36).

Gilliland-Swetland, Anne J. *Introduction to metadata: Setting the stage*. 2000. http://ptarpp2.uitm.edu.my/ptarpprack/silibus/is772/SetStage.pdf. (Cit. on pp. 36, 45).

Greene, Mark and Dennis Meissner. "More Product, Less Process: Revamping Traditional Archival Processing". *American Archivist* 68.2. (Sept. 2005): 208–263. http://archivists.metapress.com/content/C741823776K65863. (Cit. on p. 37).

Keough, Brian and Mark Wolfe. "Moving the Archivist Closer to the Creator: Implementing Integrated Archival Policies for Born Digital Photography at Colleges and Universities". *Journal of Archival Organization* 10.1. DOI: 10.1080/15332748.2012.681266. (Jan. 2012): 69–83. (Cit. on p. 37).

Knight, Steve. *Securing the Future: Digital Preservation at the National Library of New Zealand*. 2008. Paper presented at the annual conference of the international group of ex libris users [IGeLU], madrid, september 8–10, 2008. http://igelu.org/wp-content/uploads/2010/10/12a_knight.pdf. (Cit. on p. 35).

Lesk, Michael. "Preface". *Digital preservation for libraries, archives, and museums*. Ed. Edward M. Corrado and Heather Lea Moulaison. Lanham: Rowman & Littlefield, 2014. (Cit. on p. 44).

Miller, Steven J. *Metadata for digital collections: a how-to-do-it manual*. How-to-do-it manuals no. 179. New York: Neal-Schuman Publishers, 2011. (Cit. on p. 36).

Moulaison, Heather Lea. "A New Cataloging Curriculum in a Time of Innovation: Exploring a Modular Approach to Online Delivery". *Cataloging & Classification Quarterly* 50.2-3. DOI: 10.1080/01639374.2011.653096. (2012): 94–109. (Cit. on p. 45).

——. "Subject access to materials in online library catalogues: discovery of Moroccan publications". *The Journal of North African Studies* 15.3. (2010): 385–397.

National Information Standards Organization (NISO). *Understanding Metadata*. Bethesda, Md.: NISO, 2004. http://www.niso.org/publications/press/UnderstandingMetadata.pdf. (Cit. on p. 35).

Open Archives Initiative Protocol for Metadata Harvesting. 2001. http://www.openarchives.org/OAI/1.1/openarchivesprotocol.htm. (Cit. on p. 42).

Open Archives Initiative Protocol for Metadata Harvesting. 2008. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm. (Cit. on p. 43).

Park, Jung-Ran. "Metadata Quality in Digital Repositories: A Survey of the Current State of the Art". *Cataloging & Classification Quarterly* 47.3-4. DOI: 10.1080/01639370902737240. (2009): 213–228. (Cit. on p. 44).

Raymond, Matt. *My Friend Flickr: A Match Made in Photo Heaven*. 2008. http://blogs.loc.gov/loc/2008/01/my-friend-flickr-a-match-made-in-photo-heaven. (Cit. on p. 37).

Walsh, Maureen P. "Automated reuse of embedded image metadata for the Knowledge Bank". *Digital Scholarship @ The Libraries*. (). (Cit. on p. 38).

——. "Repurposing embedded image metadata for DSpace batch loading (XMP to CSV to DSpace Dublin Core)". *Metadata and More*. (2011). <http://www.mpwalshmetadata.org/2011/10/repurposing-embedded-image-metadata-for.html>. (Cit. on p. 38).

Zavalina, Oksana L. "Complementarity in Subject Metadata in Large-Scale Digital Libraries: A Comparative Analysis". *Cataloging & Classification Quarterly* 52.1. DOI: 10.1080/01639374.2013.848316. (Jan. 2014): 77–89. (Cit. on p. 41).

EDWARD M. CORRADO, Binghamton University Libraries.
ecorrado@binghamton.edu

RACHEL JAFFE, University of California, Santa Cruz.
jaffer@ucsc.edu

Corrado, E.M., R. Jaffe. «Transforming and enhancing metadata for enduser discovery: a case study». *JLIS.it*. Vol. 5, n. 2 (Luglio/July 2014): Art: #10069. DOI: 10.4403/jlis.it-10069. Web.

ABSTRACT: This paper describes the process developed by Binghamton University Libraries to extract embedded metadata from over 350,000 digital photographs and to transform this metadata into descriptive metadata for use in the Libraries' digital preservation system. Most of these images depict campus events, such as Homecoming, Commencement, etc. that are of historical and immediate social value to the campus community. However, owing to volume of photographs, as well as to budgetary and other constraints, it is not possible to have library staff inspect the photographs and create a complete descriptive metadata record for each, so we needed to explore different options.

KEYWORDS: Digital preservation; Metadata; Photographs.

ACKNOWLEDGMENT: Presented at FSR 2014 Conference, Rome, 27-28 February, 2014.

Submitted: 2014-04-15
Accepted: 2014-05-18
Published: 2014-07-01